

THE RELIABILITY OF QUARTERLY MARKS
IN THE SEVENTH GRADE OF JUNIOR
HIGH SCHOOL, TOGETHER WITH
THE VALUE OF CERTAIN
STANDARD TESTS IN
PREDICTING THEM

BY
J. CAREY TAYLOR

A DISSERTATION

submitted to the Advisory Board of the School of Higher Studies in Education of The
Johns Hopkins University in conformity with the requirements for the degree of
Doctor of Education

1930

BALTIMORE
THE JOHNS HOPKINS PRESS
1931

THE RELIABILITY OF QUARTERLY MARKS
IN THE SEVENTH GRADE OF JUNIOR
HIGH SCHOOL, TOGETHER WITH
THE VALUE OF CERTAIN
STANDARD TESTS IN
PREDICTING THEM



Digitized by the Internet Archive
in 2026

THE RELIABILITY OF QUARTERLY MARKS
IN THE SEVENTH GRADE OF JUNIOR
HIGH SCHOOL, TOGETHER WITH
THE VALUE OF CERTAIN
STANDARD TESTS IN
PREDICTING THEM

BY
J. CAREY TAYLOR

A DISSERTATION

submitted to the Advisory Board of the School of Higher Studies in Education of The
Johns Hopkins University in conformity with the requirements for the degree of
Doctor of Education

1930

BALTIMORE
THE JOHNS HOPKINS PRESS
1931

COPYRIGHT 1931 BY
THE JOHNS HOPKINS PRESS
BALTIMORE, MD.



Composed and printed by
THE LORD BALTIMORE PRESS
BALTIMORE, MD., U. S. A.

ACKNOWLEDGMENT

While it is not possible to mention specifically all who have cooperated in the making of this study, the writer wishes to acknowledge his special indebtedness to the several teachers of the Edgar Allan Poe Junior High School, whose unfailing courtesy, cooperation, and patience made possible the collection of data; to the Baltimore Junior High School principals who so generously placed their school records at his disposal; to Dr. Florence E. Bamberger and Dr. David E. Weglein, who were ever willing to offer helpful counsel; and especially to Dr. Fowler D. Brooks for his guidance and continued stimulation which have in a large measure given to the study whatever merit it may possess.

CONTENTS

CHAPTER	PAGE
I. THE PROBLEM, MATERIAL, AND METHOD.....	1
II. THE VALUE OF THE OTIS GROUP TEST AND STANFORD ACHIEVEMENT TESTS IN READING AND ARITHMETIC IN PREDICTING SCHOLARSHIP IN THE FIRST YEAR OF THE JUNIOR HIGH SCHOOL.	5
III. SUCCESS IN THE JUNIOR HIGH SCHOOL OF ENTERING STUDENTS	9
A. At Certain Levels on the Otis Group Intelligence Test.	
B. At Certain Levels on the Stanford Achievement Tests in Reading and Arithmetic.	
IV. THE RELIABILITY OF TEACHERS' MARKS, THE CRITERION OF SEVENTH GRADE SUCCESS.....	24
V. CONCLUSION	49
VI. BIBLIOGRAPHY	51
INDEX	53
VITA	55

TABLES

TABLE	PAGE
1. The Correlation of High School Marks and Group Intelligence Test Results (Collected by Jordan).....	5
2. Correlations Showing the Value of Certain Achievement and Intelligence Tests for Predicting Scholarship in the First Year of the Junior High School (Brooks).....	6
3. Simple and Multiple Correlations Showing the Relative Predictive Values of the Otis Group, Stanford Reading, and Stanford Arithmetic Tests—Edgar Allan Poe Junior High School	7
4. Simple and Multiple Correlations Showing the Relative Predictive Values of the Otis Group, Stanford Reading, and Stanford Arithmetic Tests—Twelve Junior High Schools..	8
5. Distribution of 7B Pupils in the Edgar Allan Poe Junior High School in the Otis Group, Stanford Reading, and Stanford Arithmetic Tests.....	10
6. Distribution of 7B Pupils in Twelve Junior High Schools in the Otis Group, Stanford Reading, and Stanford Arithmetic Tests.	11
7. Success in English, Mathematics, Geography, History, and Average of 7B Pupils Who Scored below 26 in the Otis Group Test.....	12
8. Success in English, Mathematics, Geography, History, and Average of 7B Pupils falling within the Third and Fourth Grade Levels of the Stanford Reading Test.....	12
9. Success in English, Mathematics, Geography, History, and Average of 7B Pupils falling within the Third and Fourth Grade Levels of the Stanford Arithmetic Test.....	13
10. Success in English, Mathematics, Geography, History, and Average of 7B Pupils making Scores between 26 and 33 in the Otis Group Test.....	14
11. Success in English, Mathematics, Geography, History and Average of 7B Pupils falling within the Fifth Grade Level in the Stanford Reading Test.....	14
12. Success in English, Mathematics, Geography, History, and Average of 7B Pupils falling within the Fifth Grade Level in the Stanford Arithmetic Test.....	14
13. Success in English, Mathematics, Geography, History, and Average of 7B Pupils making Scores between 34 and 40 in the Otis Group Test.....	15
14. Success in English, Mathematics, Geography, History, and Average of 7B Pupils falling within the Sixth Grade Level in the Stanford Reading Test.....	15
15. Success in English, Mathematics, Geography, History, and Average of 7B Pupils falling within the Sixth Grade Level in the Stanford Arithmetic Test.....	16
16. Showing the Coefficients of Reliability of the Various Elements and of the Quarterly Marks given by Teacher "A" to 78 Pupils in 7B English.....	31
17. Showing the Coefficients of Reliability of the Quarterly English Marks, together with the Reliability of Each Factor of which they are composed, given by Teacher "B" to 59 7B Pupils	33

TABLE	PAGE
18. Showing the Coefficients of Reliability of Teacher "C's" Quarterly 7B English Marks, together with the Reliability of Each Mark of which they are composed.....	34
19. Showing the Reliability of Teacher "D's" Quarterly 7B English Marks, together with the Reliability of Each Mark of which they are composed.....	35
20. Showing the Coefficients of Reliability of Teacher "E's" Quarterly 7B English Marks, together with the Reliability of Each Mark of which they are composed.....	37
21. Showing the Coefficients of Reliability of Teacher "F's" Quarterly 7B Marks in Mathematics, together with the Reliability of Each Mark of which they are composed.....	39
22. Showing the Coefficients of Reliability of Teacher "G's" Quarterly 7B Marks in Mathematics, together with the Reliability of Each Mark of which they are composed.....	40
23. Showing the Coefficient Reliability of Teacher "H's" Quarterly 7B Marks in Geography, together with the Reliability of Each Mark of which they are composed.....	41
24. Showing the Coefficients of Reliability of Teacher "I's" Quarterly 7B Marks in History, together with the Reliability of Each Mark of which they are composed.....	43
25. Showing the Coefficients of Reliability of Teacher "J's" Quarterly 7B Marks in History, together with the Reliability of Each Mark of which they are composed.....	44
26. Showing the Coefficients of Reliability of Teacher "K's" Quarterly 7B Marks in History, together with the Reliability of Each Mark of which they are composed.....	45
27. Showing the Average Reliability of the Quarterly Marks for Each Teacher, for Each Subject, and for All Teachers....	46
28. Showing the Correlation between the Scores made on the Otis Group Test, Stanford and Sangren-Woody Reading Tests, Stanford Arithmetic Test, and the Average of Marks for September-February 1928-1929, and September-February 1929-1930.	47

FIGURES

FIGURE		PAGE
1.	Showing the Distribution of Averages Below 60 and Above 80 in the various Grade Levels of the Otis Group Test at Edgar Allan Poe Junior High School.....	18
2.	Showing the Distribution of Averages Below 60 and Above 80 in the various Grade Levels of the Otis Group Test in Twelve Junior High Schools.....	19
3.	Showing the Distribution of Averages Below 60 and Above 80 in the various Grade Levels of the Stanford Reading Test at Edgar Allan Poe Junior High School.....	20
4.	Showing the Distribution of Averages Below 60 and Above 80 in the various Grade Levels of the Stanford Reading Test in Twelve Junior High Schools.....	21
5.	Showing the Distribution of Averages Below 60 and Above 80 in the various Grade Levels of the Stanford Arithmetic Test at Edgar Allan Poe Junior High School.....	22
6.	Showing the Distribution of Averages Below 60 and Above 80 in the various Grade Levels of the Stanford Arithmetic Test in Twelve Junior High Schools.....	23

CHAPTER I

THE PROBLEM, MATERIAL, AND METHOD

THE PROBLEM

In most progressive high schools large enough to permit classification of pupils, such classification is generally based upon one or more of the following: average marks received in the grade immediately preceding, scores made in intelligence tests, scores made in achievement tests, teachers' estimates of the pupils' intelligence, industry, and the like, and chronological age. As far as is known, the average mark received in preceding grades is the best single basis for predicting scholarship. Next in order of predictive value are teachers' estimates, intelligence scores, achievement scores, and chronological age.¹

In the junior high schools of Baltimore, Maryland, it is customary to administer the Otis Group Test of Mental Ability, the Stanford Reading Test, and the Stanford Arithmetic Test, or substitutes, to all entering seventh grade pupils. While these are given partly for diagnostic purposes, the results are also taken into consideration when the pupils are classified and placed in accelerate, normal, or slow classes. In most schools greater weight is given to the general rating of the previous semester's work, but the intelligence and achievement test scores are frequently carefully studied with the hope of being able to predict how successful the pupils concerned will be in the various major subjects of the junior high school.

There is a prevalent belief that while a high intelligence is predictive of general success, a high reading score indicates considerable ability in English and perhaps in other subjects, and a high score in arithmetic is nearly infallible as a prediction of success in mathematics. One junior high school is at the time of this writing grouping the seventh grade pupils according to the scores made in these two achievement tests,

¹ Fowler D. Brooks, *Psychology of Adolescence*, 1929, page 558.

so that the poor readers may be in one class and the poor arithmetic students in another.² In most schools, however, there is a feeling of uncertainty about just how much weight should be given to these tests when classifying pupils.

Therefore, it is with the predictive value of the Otis Intelligence Test, the Stanford Reading Test, and the Stanford Arithmetic Test that the first part of this study is concerned. Does a high reading score mean success in English, and does a high arithmetic quotient signify success in the type of mathematics taught in the seventh grade of the Baltimore junior high schools?

When we speak of tests as predictive of the success of pupils in certain subjects, we consider that success as indicated by the teachers' marks. In other words, we are concerned with the correlation between the test scores and term marks received by the individual, regardless of how accurate a measure of his ability these marks may be. As far as the pupil's school life is concerned, his progress is wholly determined by the teachers' marks, whether those marks are based, solely on the teachers' judgments, on the results of tests, on the grades received on written compositions, on some objective measures, or various combinations of these or similar data. Much has been said about the variability of teachers' marks. Experiments have shown that teachers' marks vary.^{3, 4} There seems to be too much of the personal element involved to keep them constant. A single mark given by a teacher may be unreliable, but since the reliability of an average increases when the number of marks becomes greater, is a teacher's term mark necessarily a very unreliable measure of the pupil's achievement? With the realization both of the importance and probable inaccuracy of the teacher's term mark, the second part of this study undertakes to determine the degree of reliability that may be expected in the term marks of seventh grade teachers in general.

² Junior High School No. 40, Baltimore, Md.

³ Starch and Elliott—Reliability of the Grading of High School Work in English—School Review, Vol. 20 (1912), pages 442-457.

⁴ W. F. Dearborn—School and University Grades—Bulletin of the University of Wisconsin, No. 368 (1910), page 57.

THE MATERIAL AND METHOD

In the first part of the study, which undertakes to determine the predictive value of scores made in the Otis Group Test and Stanford Achievement Tests, two groups of 7B pupils were used. The first consisted of 204 pupils who entered the Edgar Allan Poe Junior High School in September, 1928. The second was a sample group of 350 pupils who entered the 12 white junior high schools of Baltimore at the same time. The author endeavored to make this a normal group, following in every respect the shape of curves formed by the chronological ages and intelligence scores of all the 7B pupils in the city. In the 12 junior high schools there were 3496 7B Students. The 350 were chosen by taking 10% of the enrollment of each school. Furthermore, the pupils from each school were so selected that with respect both to intelligence and chronological age, the sample group was typical of the whole class. Since this was done in each school, the 12 sections, when placed together, formed a typical curve in these two respects.

In September, 1928, as part of the city-wide survey of pupils entering the junior high schools, the following tests were given:

- Otis Self Administering Group Test of Mental Ability
- Intermediate Examinations, Form A
- Stanford Achievement Test
- Reading—Parts 1, 2, and 3
- Arithmetic—Parts 4, and 5

For the pupils within the one school, and for the representative group, the results of these tests, together with the semester marks at the end of the 7B grade, furnish the material from which it is possible to state the predictive value of the tests concerned.

In the second part of the study an effort is made to determine the reliability of teachers' marks. The marks in English, mathematics, geography, and history of a group of 226 seventh grade pupils who entered the Edgar Allan Poe Junior High School in September, 1929, provided the material for the investigation.

When these pupils entered the junior high school they were given the Otis Self-Administering Group Test of Mental Ability, the Sangren-Woody Test in Reading, Parts 1, 2, and 3, and the Stanford Test in Arithmetic, Parts 4 and 5. They were divided into seven classes, the classification being based on the teacher's estimate of the work of the sixth grade, the mental age, intelligence quotient, and the scores made in the achievement tests. 7B¹ was considered as the brightest class, 7B² next, 7B⁷ the slowest, and so on. From the time these classes began work in September, every teacher to whom they were assigned for English, mathematics, geography, or history was requested to teach and mark the work of these pupils as he had always done, but to give the author at the end of each quarter a complete list of all the marks from which the pupils' quarterly marks were determined, together with all written compositions, exercises, tests, and papers of any kind whose marking influenced in any way the pupils' average for the quarter.

By means of self-correlation of tests, remarking of compositions, and dividing the sums of several recitation marks into chance halves and then correlating, coefficients of reliability of all of the many factors involved were determined. From these several coefficients of reliability, grouped and weighted as were the teachers' marks before averaging, it was possible to determine the reliability of the teachers' quarterly marks.

CHAPTER II

THE VALUE OF THE OTIS GROUP TEST AND STANFORD ACHIEVEMENT TESTS IN READING AND ARITHMETIC IN PREDICTING SUCCESS IN THE FIRST YEAR OF JUNIOR HIGH SCHOOL

As was mentioned in the introduction, intelligence and achievement tests play an important part in the classification of school children, teachers believing them to predict with a fair degree of accuracy the school achievement of the pupils

TABLE 1

THE CORRELATION OF HIGH SCHOOL MARKS AND GROUP INTELLIGENCE
TEST RESULTS (COLLECTED BY JORDAN)

	All subjects	English	Mathematics	General Science	History
Otis	.450	.466	.430	.502	.262
Army	.476	.472	.511	.456	.319
Miller	.476	.456	.456	.592	.168
Terman	.492	.436	.466	.636	.408

(This means that the correlation between the scores made in the Otis Group Test and the average of the term marks in all subjects is .450, etc.)

as measured by the teachers' marks. Several investigations justify their use. Jordan, University of Iowa,¹ in correlating intelligence with grades of 67 high school pupils, found that the Otis Group Test correlated .460 with English marks, .430 with mathematics, .502 with science, .262 with history, and .450 with all subjects. The Army Alpha, Terman, and Miller Tests gave similar predictions. (See Table 1.) Brooks,² in a study made with 93 junior high school pupils, obtained similar results when the scores made in various intelligence and achievement tests were correlated with marks in grade VII.

¹ A. M. Jordan—Correlation of Four Intelligence Tests with Grades—Journal of Educational Psychology (1922), pages 419-429.

² Fowler D. Brooks—Sectioning Junior High School Pupils by Tests and School Marks—Journal of Educational Research, Vol. 12 (1925), pages 359-369.

(See Table 2.) Haddock³ found that eighth grade students who made an E. Q. of 105 or over in the Stanford Achievement Test, on the whole receive high marks in the ninth grade. Other investigations⁴ show correlations closely enough similar to the ones presented to indicate that they are typical of the relationship existing between scores made in intelligence and achievement tests and school marks.

TABLE 2
CORRELATIONS SHOWING THE VALUE OF CERTAIN ACHIEVEMENT AND
INTELLIGENCE TESTS FOR PREDICTING SCHOLARSHIP IN THE
FIRST YEAR OF THE JUNIOR HIGH SCHOOL (BROOKS)

		Marks in Grade VII
Thorndike-McCall Reading		.511
Holly Sentence Vocabulary		.462
Kelly-Trabue Language Completion Alpha		.296
Group Vocabulary Test (Binet List)		.400
Woody-McCall Arithmetic		.342
Dearborn Intelligence	C	.476
Dearborn Intelligence	D	.474
Haggerty Delta Intelligence	2	.450
Illinois Intelligence	A	.412
Miller Intelligence	A	.501
National Intelligence	A	.496
Otis Self Administering	A	.541
Pintner, Non-Language		.304
Terman	A	.401

Granting then that these intelligence and achievement tests have some predictive value, the purpose of this study is to decide just how much weight can be placed in a single administering of each of the three now in use in the Junior High Schools of Baltimore: namely, the Otis Group Test of Mental Ability, the Stanford Reading Test, and the Stanford Arithmetic Test.

In September, 1928, these tests were given to 3496 pupils who entered the low seventh grade in the Baltimore schools. Two experimental groups were selected. The first was composed of 204 pupils, the entering class at the Edgar Allan Poe Junior High School; the second of 350 pupils, selected, as

³ Nellie P. Haddock—The Use of the Stanford Achievement Test to Indicate High School Success—School and Society (January 22, 1927), pages 114-116.

⁴ See Bibliography—Book, Ohlson, Proctor, Rector, Senour, and Unzicker.

explained in the Introduction, from the 3496 boys and girls in the 12 junior high schools. Because of withdrawals, the groups were finally reduced to 166 and 316, respectively. At the end of the first semester the scores made in each test were correlated with the teachers' marks in English, mathematics, geography, history, and the average of the four. Their multiple correlations were used to see if a combination of any two of the three tests would predict more highly than any single test.

A study of Table 3 shows that within the one school the Otis Group, the Stanford Arithmetic Test, and the combina-

TABLE 3

SIMPLE AND MULTIPLE CORRELATIONS SHOWING THE RELATIVE PREDICTIVE VALUES OF THE OTIS GROUP, STANFORD READING, AND STANFORD ARITHMETIC TESTS—EDGAR ALLAN POE JUNIOR HIGH SCHOOL

	Eng.	Math.	Geog.	N=166 Hist.	Avg
Otis Group	.341	.255	.438	.477	.474
Stanford Reading	.171	.223	.240	.347	.307
Stanford Arithmetic	.342	.472	.481	.433	.556
Otis Group and Stanford Reading	.209	.226	.282	.360	.331
Otis Group and Stanford Arithmetic	.342	.497	.482	.513	.559
Stanford Reading and Stanford Arithmetic	.179	.238	.254	.349	.317
Otis, Reading and Arithmetic	.179	.240	.254	.347	.317

tion of the two predict equally well success in English. The Stanford Arithmetic and the Otis Group predict success in arithmetic better than the former test when taken alone. Geography success is best predicted by either the arithmetic test alone, or by combining it with the Otis. Success in history and all four subjects taken together is best foretold by the Otis and Stanford Arithmetic combined. The predictive value of the Stanford Arithmetic Test is outstanding in all subjects except history. The fact that these pupils tested slightly above the average in arithmetic, and slightly below in reading and intelligence, may be the cause for the position

the arithmetic test assumes. All of these correlations, however, are too low for good prediction. According to Kelly,

when "r" is	20	the error of prediction is	.9798
" " " 30	" " " "	" " " "	.9539
" " " 40	" " " "	" " " "	.9165
" " " 50	" " " "	" " " "	.8660
" " " 60	" " " "	" " " "	.8000
" " " 70	" " " "	" " " "	.7141
" " " 80	" " " "	" " " "	.6000
" " " 90	" " " "	" " " "	.4359

Since the coefficient of correlation must be .85 before the error is half way between a guess and perfect prediction, many of these coefficients are of little significance.

In Table 4 it will be noticed that the Otis Group Test predicts much more highly for each subject and for the average

TABLE 4
SIMPLE AND MULTIPLE CORRELATIONS SHOWING THE RELATIVE PREDICTIVE VALUES OF THE OTIS GROUP, STANFORD READING, AND STANFORD ARITHMETIC TESTS—TWELVE JUNIOR HIGH SCHOOLS

	N=316				
	Eng.	Math.	Geog.	Hist.	Avg.
Otis Group	.479	.445	.448	.477	.541
Stanford Reading	.394	.356	.360	.406	.473
Stanford Arithmetic	.278	.391	.389	.364	.446
Otis Group and Stanford Reading	.405	.356	.374	.413	.479
Otis Group and Stanford Arithmetic	.312	.393	.391	.373	.451
Stanford Reading and Stanford Arithmetic	.399	.357	.362	.407	.473
Otis, Reading, and Arithmetic	.399	.358	.361	.407	.473

of all subjects than does either of the other tests or any combinations of them. The tests in reading and arithmetic share a secondary position, the reading test having more weight in English and history, and the arithmetic test in mathematics and geography.

Since the second group is not only more representative, but also nearly twice as large as the former, its results should be considered more seriously. In so far as this sample is typical of all the seventh grades of Baltimore, there is little reason for considering the scores made in the Stanford Reading and Arithmetic Tests when we are endeavoring to predict the success of pupils in English, mathematics, geography, history, or the average of these four.

CHAPTER III

SUCCESS IN THE JUNIOR HIGH SCHOOL OF ENTERING STUDENTS FALLING WITHIN CERTAIN LEVELS OF THE OTIS GROUP AND STANFORD ACHIEVEMENT TESTS

The correlations of the preceding chapter show the general proportionality existing between the scores made by seventh grade pupils in the Otis Group and Stanford Achievement Tests and their marks in certain seventh grade subjects. From this the comparative value of the tests for predictive purposes can be determined. Such a procedure, however, does not enable one to predict the chances of seventh grade success of an individual who makes a certain test score. In this chapter the term marks of pupils falling at certain levels on these tests will be reviewed for the purpose of making a prediction of the seventh grade success of such pupils.

Tables 5 and 6 show the entire distribution of test scores and term marks of the 7B pupils in the Edgar Allan Poe Junior High School and the group selected from 12 junior high schools. In each of the Stanford Achievement tests the scores are grouped according to grade levels determined by the author. The grade levels of the Otis Group Test were determined by the Bureau of Research of the Baltimore Department of Education. In this chapter these levels are grouped as follows:

1. Third and fourth grade levels
2. Fifth grade
3. Sixth grade
4. Seventh grade
5. Eighth grade
6. Ninth grade and above

Tables 5 and 6 give the entire distribution of the scores made in the Otis Group and Stanford Arithmetic Tests, together with the semester marks in English, mathematics, geography, history, and the average of these.

TABLE 5
DISTRIBUTION OF 7B PUPILS IN EDGAR ALLAN POE JUNIOR HIGH SCHOOL IN THE OTIS GROUP, STANFORD
READING, AND STANFORD ARITHMETIC TESTS

Test Scores	Number	English					Mathematics					Geography					History					Average					
		65-0	66-9	70-4	68-8	001-06	65-0	66-9	70-4	68-8	001-06	65-0	66-9	70-4	68-8	001-06	65-0	66-9	70-4	68-8	001-06	65-0	66-9	70-4	68-8	001-06	
Otis Group																											
Third Grade Level	3	1	2	..	..	..	1	..	1	..	1	1	2	..	..	..	..	..	..	..	..	1	2	..	..	..	..
Fourth "	12	4	3	4	1	..	5	2	..	..	..	4	5	1	2	..	2	8	1	1	..	6	2	4	..	..	
Fifth "	31	5	9	11	5	1	4	13	8	5	1	3	14	11	3	..	1	24	5	..	1	5	12	14	..	..	
Sixth "	34-40	41	5	19	12	4	10	18	9	4	..	2	13	16	5	..	3	20	11	6	1	9	17	14	1	..	
Seventh "	41-47	39	2	16	16	5	4	16	14	5	..	2	11	21	5	..	2	14	14	7	2	3	15	21	..	..	
Eighth "	48-51	14	..	6	7	1	1	6	4	2	1	..	3	7	3	1	1	3	6	2	2	1	4	7	2	..	
Ninth "	52-55	11	..	3	6	2	..	2	4	4	1	..	2	3	6	..	..	2	1	6	2	..	2	5	4	..	
Tenth "	56-59	13	..	3	7	3	..	..	4	3	6	..	..	1	6	4	..	..	2	5	1	..	1	9	1	..	
Eleventh "	60-75	2	..	..	..	..	..	..	1	1	..	..	..	2	..	..	..	1	..	1	..	..	..	1	1	..	
	166	17	61	64	22	2	25	61	47	29	4	17	51	67	28	3	11	75	43	28	9	25	55	75	9	2	
Stanford Reading																											
Third Grade Level	0-81	1	1	..	..	..	1	..	..	..	..	1	..	..	..	..	1	..	..	..	..	1	..	..	..	..	
Fourth "	82-116	18	5	6	6	1	..	5	5	4	2	2	5	5	3	..	2	11	4	..	1	7	4	6	1	..	
Fifth "	117-143	55	5	18	20	10	2	9	23	14	8	1	4	18	25	7	1	4	28	13	7	3	7	27	19	2	
Sixth "	144-163	40	6	13	17	4	..	5	17	13	5	..	5	15	13	7	..	2	19	12	6	1	8	10	20	2	
Seventh "	164-181	26	..	13	11	2	..	3	9	9	5	..	1	7	15	3	..	1	10	7	6	2	1	10	14	1	
Eighth "	182-199	16	..	5	7	4	..	1	3	4	7	1	..	5	6	4	1	1	6	5	3	1	1	3	9	2	
Ninth "	200-216	9	..	6	2	1	..	1	4	3	1	..	1	1	3	3	1	..	1	2	5	1	..	1	6	1	
Tenth "	217-275	1	..	..	..	..	..	..	..	..	1	..	..	..	..	1	..	..	..	..	..	..	..	1	..	..	
	166	17	61	64	22	2	25	61	47	29	4	17	51	67	28	3	11	75	43	28	9	25	55	75	9	2	
Stanford Arithmetic																											
Third Grade Level	0-96	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	
Fourth "	97-130	12	7	3	2	..	..	7	4	1	..	..	5	6	1	..	..	5	7	..	..	8	3	1	..	..	
Fifth "	131-163	47	7	23	13	3	1	13	20	7	9	1	8	18	17	4	..	2	34	7	3	1	10	23	13	1	
Sixth "	164-188	56	2	21	22	10	1	5	23	19	9	1	4	16	23	12	1	2	37	18	10	4	5	18	31	2	
Seventh "	189-211	32	..	10	17	5	..	..	11	16	4	1	5	20	6	1	1	2	7	12	8	3	1	8	19	3	
Eighth "	212-233	16	1	4	8	3	..	..	4	4	7	1	..	6	6	3	1	1	3	6	5	1	1	3	10	1	
Ninth "	234-245	2	1	..	..	1	..	..	..	..	..	..	..	..	..	1	..	..	2	1	..	..	..	..	1	1	
Tenth "	246-348	2	..	..	..	..	..	..	..	..	2	..	..	..	..	2	..	..	1	..	..	..	..	..	1	..	
	166	17	61	64	22	2	25	61	47	29	4	17	51	67	28	3	11	75	43	28	9	25	55	75	9	2	

TABLE 6

DISTRIBUTION OF 7B PUPILS IN TWELVE JUNIOR HIGH SCHOOLS IN THE OTIS GROUP, STANFORD READING,
AND STANFORD ARITHMETIC TESTS

Otis Group	Test Scores	Number	English				Mathematics				Geography				History				Average			
			001-06	07-12	13-18	19-24	25-30	31-36	37-42	43-48	49-54	55-60	61-66	67-72	73-78	79-84	85-90	91-96	97-100	101-106	107-112	113-118
Third Grade	0-15	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Fourth	16-25	21	9	7	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Fifth	26-33	53	15	24	10	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sixth	34-40	66	7	25	22	11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Seventh	41-47	74	7	20	23	22	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eighth	48-51	35	1	9	11	10	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ninth	52-55	28	2	2	13	11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tenth	56-59	22	1	1	8	9	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eleventh	60-75	15	1	1	3	7	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		316	42	90	94	77	13	48	107	72	64	25	25	76	96	89	30	23	128	76	61	28
Stanford Reading	0-81	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third Grade	82-116	26	6	11	5	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Fourth	117-143	69	17	25	15	12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Fifth	144-163	67	10	24	21	10	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sixth	164-181	70	4	17	27	17	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Seventh	182-199	42	3	7	14	17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eighth	200-216	23	1	5	8	7	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ninth	217-275	17	1	1	4	10	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tenth		316	42	90	94	77	13	48	107	72	64	25	25	76	96	89	30	23	128	76	61	28
Stanford Arithmetic	0-96	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third Grade	97-130	83	6	9	5	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Fourth	131-163	22	17	20	18	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Fifth	164-188	110	11	31	41	21	6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sixth	189-211	56	5	14	18	16	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Seventh	212-233	35	2	6	7	18	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Eighth	234-248	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ninth		316	42	90	94	77	13	48	107	72	64	25	25	76	96	89	30	23	128	76	61	28
Tenth		316	42	90	94	77	13	48	107	72	64	25	25	76	96	89	30	23	128	76	61	28

We shall first consider Group 1. Table 7 shows that in the Edgar Allan Poe Junior High School there were 15 pupils who made a score of less than 26 in the Otis Group Intelligence Test. Approximately half of this group failed in each subject, and all but one had a general average of less than 60. Of the

TABLE 7

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND AVERAGE OF 7B PUPILS WHO SCORED BELOW 26 IN THE OTIS GROUP TEST

Subject	Edgar Allan Poe Junior High School N=15		Twelve Junior High Schools N=24	
	Passing	Failing	Passing	Failing
English	10	5	14	10
Mathematics	9	6	18	6
Geography	10	5	20	4
History	11	4	20	4
Average	8	7	17	7

24 pupils in the 12 junior high schools, there were 10 failures in English and 6 in mathematics, but geography and history apparently presented little difficulty. Of the 24, seven had a failing average.

In the Stanford Reading Test 19 pupils fell within the third and fourth grade levels in the one school, and 28 within

TABLE 8

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND AVERAGE OF 7B PUPILS FALLING WITHIN THE THIRD AND FOURTH GRADE LEVELS OF THE STANFORD READING TEST

Subject	Edgar Allan Poe Junior High School N=19		Twelve Junior High Schools N=28	
	Passing	Failing	Passing	Failing
English	13	6	21	7
Mathematics	13	6	19	9
Geography	13	6	22	6
History	16	3	25	3
Average	11	8	20	8

the other group. In both these groups the failures ranged between one third and one half in all subjects except history, in which there were very few marks below 60. (See Table 8.)

In the Stanford Arithmetic Test twelve pupils from the Edgar Allan Poe School fell within the third and fourth

grade levels. More than half of these failed in English and mathematics, less than half in geography and history, and two thirds in all subjects. In the group of 23 selected from other schools, approximately one fourth were unsuccessful in each subject except geography, in which all but two passed. (See Table 9.)

Even though these pupils seem to be of very low ability, the majority are able to complete the work of grade. It is true that most of these, as Tables 5 and 6 show, receive below 70 in their subjects, but a few are able to keep up with the mass of other pupils of superior ability. Their condition seems by no means a hopeless one.

TABLE 9

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND
AVERAGE OF 7B PUPILS FALLING WITHIN THE THIRD AND
FOURTH GRADE LEVELS OF THE STANFORD ARITHMETIC TEST

Subject	Edgar Allan Poe Junior High School N=12		Twelve Junior High Schools N=23	
	Passing	Failing	Passing	Failing
English	5	7	17	6
Mathematics	5	7	18	5
Geography	7	5	21	2
History	7	5	17	6
Average	4	8	18	5

Group 2, composed of pupils of the fifth grade level, is a little more successful. In the Edgar Allan Poe Junior High School 31 pupils made a score of less than 34 in the Otis Group Test. In the group selected from 12 schools, 53 fell within this level. With the exception of English in the one, the median mark for each subject in both groups fell between 60 and 69, with generally more above than below this number. (See Table 10.) Referring again to Tables 5 and 6, we find that a few pupils made marks between 80 and 90, and occasionally above.

In the Stanford Reading Test 55 pupils from the one junior high school and 69 from the selected group fall within the fifth grade level. Here again, with the exception of English in the Edgar Allan Poe School and geography in both

groups, the median mark lies between 60 and 69. In every case there are more above than below this range. (See Table 11.)

TABLE 10

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND
AVERAGE OF 7B PUPILS MAKING SCORES BETWEEN
26 AND 33 ON THE OTIS GROUP TEST

Subject	Edgar Allan Poe Junior High School N=31			Twelve Junior High Schools N=53		
	0-59	60-69	Over 69	0-59	60-69	Over 69
English	5	9	17	15	24	14
Mathematics	4	13	14	18	23	12
Geography	3	14	14	11	16	26
History	1	24	6	8	31	14
Average	5	12	14	14	25	14

TABLE 11

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND
AVERAGE OF 7B PUPILS FALLING WITHIN THE FIFTH
GRADE LEVEL ON THE STANFORD READING TEST

Subject	Edgar Allan Poe Junior High School N=55			Twelve Junior High Schools N=69		
	0-59	60-69	Over 69	0-59	60-69	Over 69
English	5	18	32	17	25	27
Mathematics	9	23	23	15	26	28
Geography	4	18	33	9	25	35
History	4	28	23	9	39	21
Average	7	27	21	14	27	28

TABLE 12

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND
AVERAGE OF 7B PUPILS FALLING WITHIN THE FIFTH
GRADE LEVEL ON THE STANFORD ARITHMETIC TEST

Subject	Edgar Allan Poe Junior High School N=56			Twelve Junior High School N=83		
	0-59	60-69	Over 69	0-59	60-69	Over 69
English	2	21	33	17	29	37
Mathematics	5	22	29	22	40	21
Geography	4	16	36	10	25	48
History	1	23	32	7	48	28
Average	5	18	33	15	41	27

A review of the scores made on the Stanford Arithmetic Test (see Table 12) shows that in the single junior high

school the median mark for every subject lies above 69. Within the larger group this is true only in geography. The mathematics marks in this group seem low, with more pupils receiving below 60 than above 69.

Apparently these pupils, all more than one full year below the grade in at least one test, have more than an even chance of successfully completing the seventh grade work. While

TABLE 13

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND
AVERAGE OF 7B PUPILS MAKING SCORES BETWEEN
34 AND 40 IN THE OTIS GROUP TEST

Subject	Edgar Allan Poe Junior High School N=41				Twelve Junior High Schools N=66			
	0-59	60-69	70-79	Over 79	0-59	60-69	70-79	Over 79
English	5	19	12	5	7	25	22	12
Mathematics	10	18	9	4	10	28	16	12
Geography	7	13	16	5	4	22	22	18
History	3	20	11	7	5	31	17	13
Average	9	17	14	1	4	33	23	6

TABLE 14

SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND
AVERAGE OF 7B PUPILS FALLING WITHIN THE SIXTH
GRADE LEVEL ON THE STANFORD READING TEST

Subject	Edgar Allan Poe Junior High School N=40				Twelve Junior High Schools N=67			
	0-59	60-69	70-79	Over 79	0-59	60-69	70-79	Over 79
English	6	13	17	4	10	24	21	12
Mathematics	5	17	13	5	12	27	9	19
Geography	5	15	13	7	6	18	23	20
History	2	19	12	7	6	24	23	14
Average	8	10	20	2	7	26	26	8

the majority of their marks lie between 60 and 69, many are considerably higher. Geography seems easier than any other subject for these children of fifth grade ability and achievement. There is no noticeable difference among the predictive values of the three tests.

Tables 13, 14, and 15 show the distribution in English, mathematics, geography, history, and general average of

Group 3, composed of pupils falling within the sixth grade levels of the Otis Group and Stanford Achievement Tests. Here a noticeable decrease in the number of failures appears, and the number of pupils getting marks of .70 or over is slightly greater than the number of those receiving below .60. The medians for the various groups fall either between .60 and .69, or between .70 and .79. In most cases the geography marks are higher than the others.

The range of scores in Table 13, in which the pupils are classified according to the scores made on the Otis Group Test, is almost identical with the range in Table 14, giving the dis-

TABLE 15
SUCCESS IN ENGLISH, MATHEMATICS, GEOGRAPHY, HISTORY, AND
AVERAGE OF 7B PUPILS FALLING WITHIN THE SIXTH
GRADE LEVEL IN THE STANFORD ARITHMETIC TEST

Subject	Edgar Allan Poe Junior High School N=56				Twelve Junior High Schools N=110			
	0-59	60-69	70-79	Over 79	0-59	60-69	70-79	Over 79
English	2	21	22	11	11	31	41	27
Mathematics	5	22	19	10	16	36	31	27
Geography	4	16	23	13	11	25	31	43
History	1	23	18	14	8	37	38	27
Average	5	18	31	2	10	29	52	19

tribution according to the grade levels as determined by the Stanford Reading Test.

The range of marks in Table 15, in which the pupils are classified according to the scores they made in the Stanford Arithmetic Test, is higher than in either of the other groups.

In Group 4, made up of those pupils who fall within the seventh grade levels of the various tests, the median mark in every case but one (the geography mark as predicted by the Stanford Reading Test at the Edgar Allan Poe School) falls between .70 and .79. A scanning of the seventh grade levels on Tables 15 and 16 make this statement so clear that additional tables would add nothing.

With each test, and in each subject, there seems to be a nearly normal curve of distribution, with few failures or very high marks.

Among the pupils in Group 5, composed of those within the eighth grade level of the three tests, there are very few failures. None of the medians of this group lie between 60 and 69, but in several subjects they fall between 80 and 90. Again there is a nearly normal curve of distribution in each case with the median a few points higher than was the case in the lower level.

In Group 6, composed of a very small number of pupils falling within the levels of the ninth grade or above, the failures in all subjects are practically eliminated and most of the marks are above 80.

A careful study of Tables 5 and 6, especially the latter, which represents a larger and more representative body of pupils, reveals that there is no noticeable difference among the predictive values of the three tests. In the case of each, the poor marks disappear and the good marks increase as the test scores improve. (See Figures 1, 2, 3, 4, 5, and 6.) However, in no case can one definitely determine by means of a test what a pupil will do in a certain subject or combination of subjects. He can predict the probable chance of success, but no more. If we interpret this statement in terms of the standard error of estimate for the various correlations found in Chapter II, we find that the per cent in reduction of the standard error of estimate ranges from 4 to 15%, with an average of about 13%.

The spread of the scores in Tables 5 and 6 enables one to understand why the correlations between the test scores and pupils' marks are relatively low.

A study of the data of Table 6 shows that few of those who were at the seventh grade level or above on these tests failed in the seventh grade, the percentages being 5, 3, and 3 for the intelligence, reading, and arithmetic tests, respectively. The other failures according to grade level were as follows: sixth grade—Otis 10%, reading 10%, arithmetic 9%; fifth grade—Otis 26%, reading 20%, arithmetic 18%; third and fourth grade levels—Otis 29%, reading 29%, arithmetic 22%.

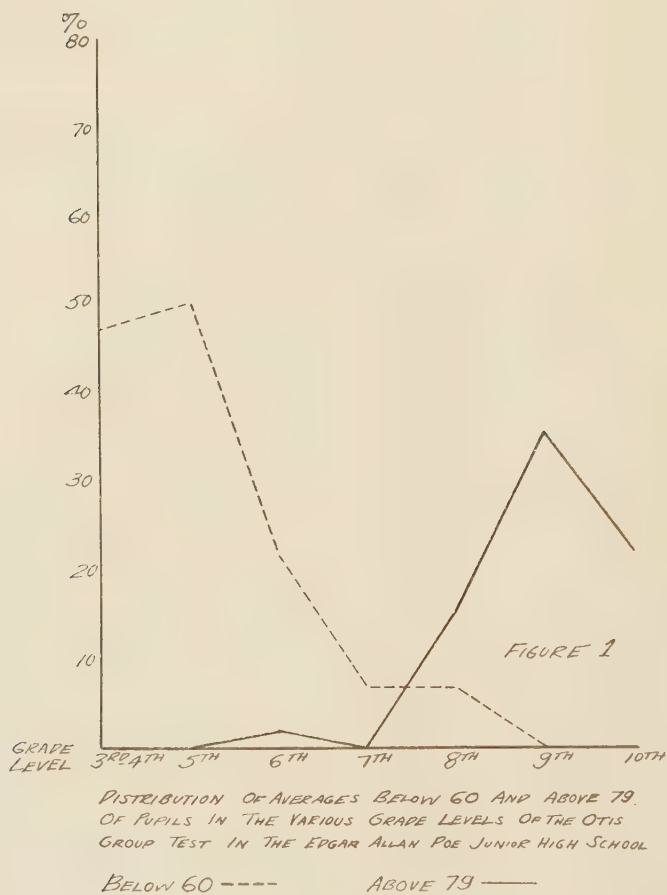


FIG. 1.



FIG. 2.

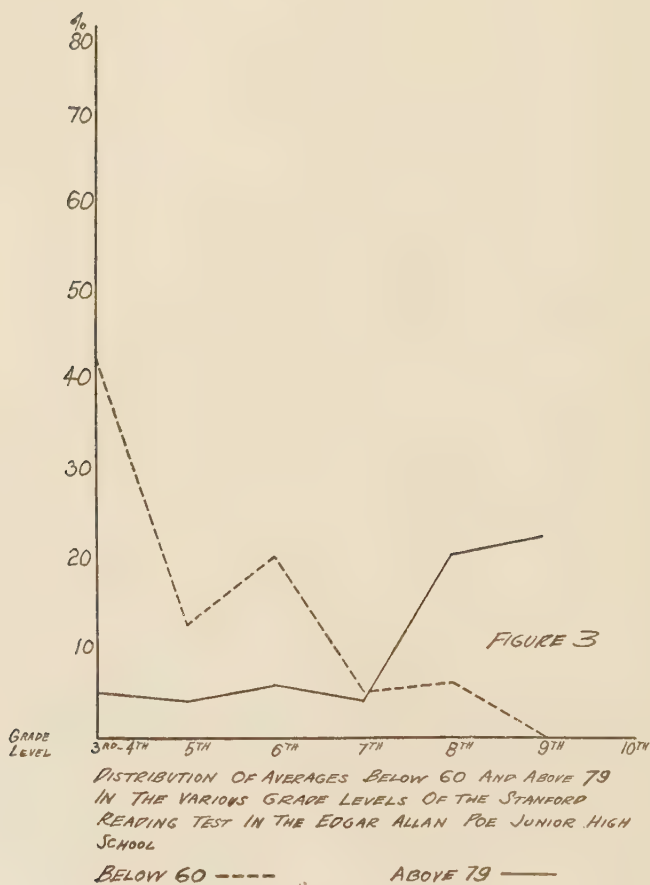
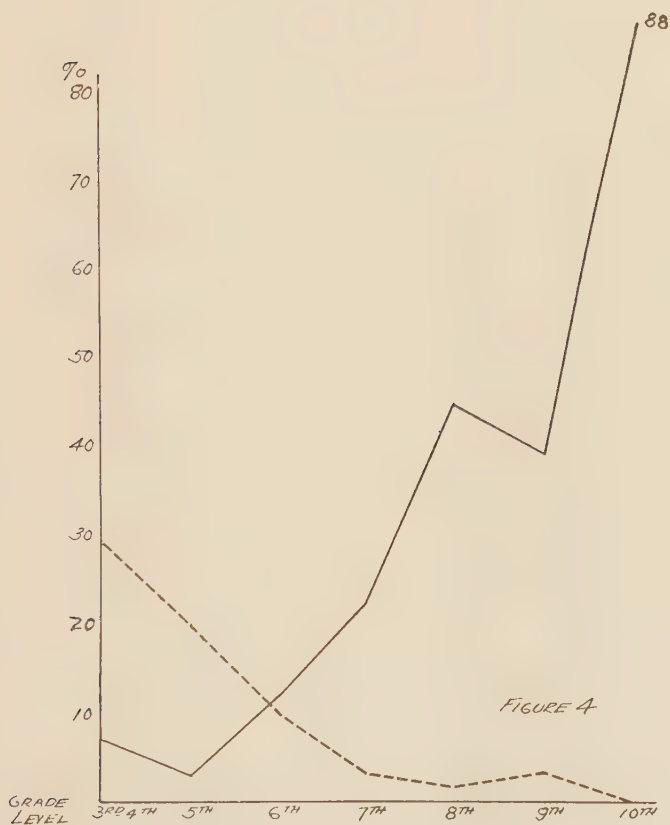


FIG. 3.



DISTRIBUTION OF AVERAGES BELOW 60 AND ABOVE 79
IN THE VARIOUS GRADE LEVELS OF THE STANFORD
READING TEST IN TWELVE JUNIOR HIGH SCHOOLS

BELOW 60 ----

ABOVE 79 ———

FIG. 4.

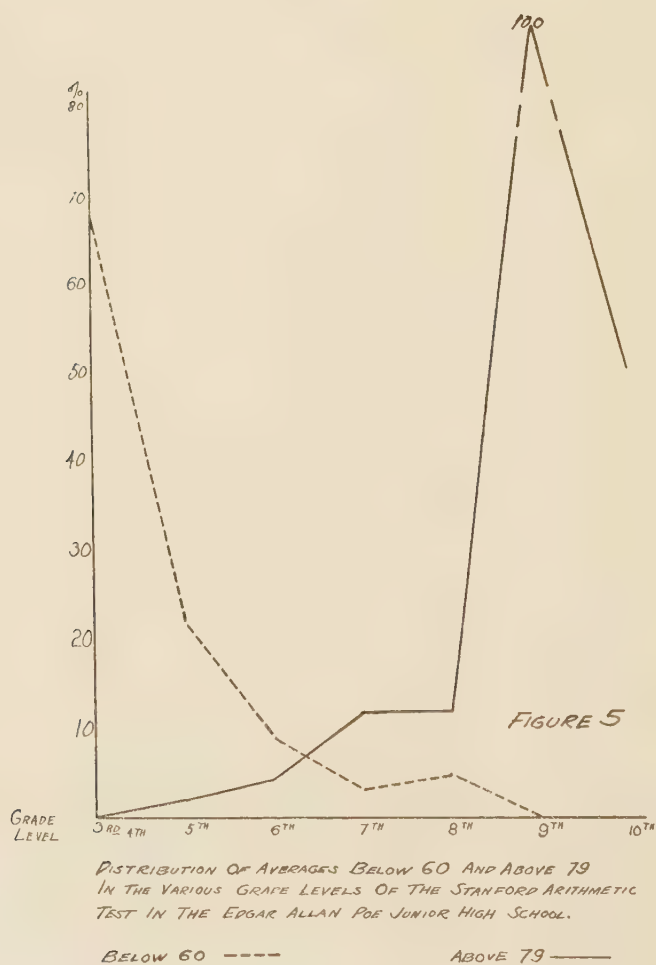


FIG. 5.

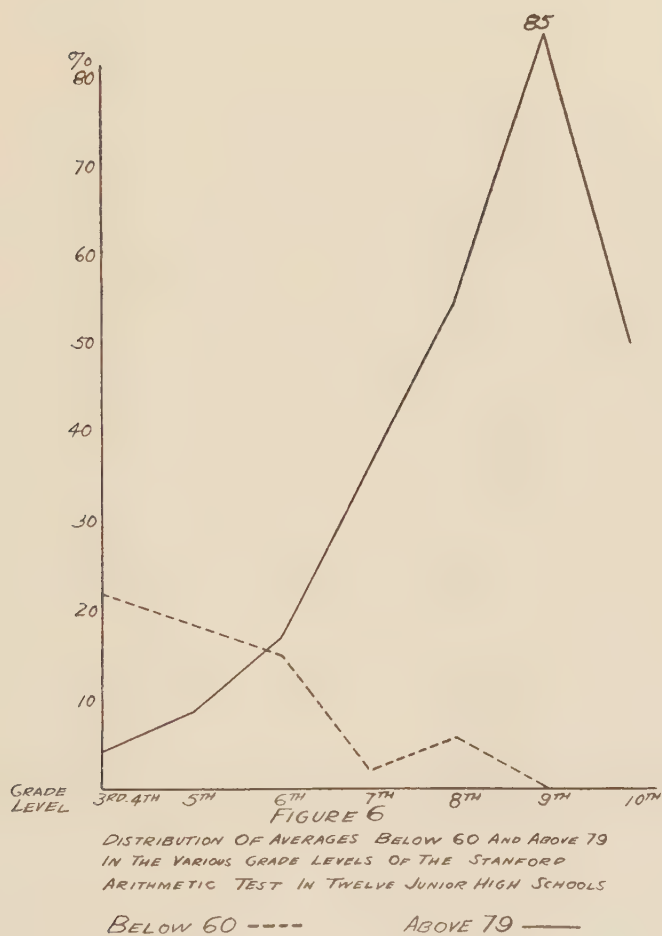


FIG. 6.

CHAPTER IV

THE RELIABILITY OF TEACHERS' MARKS, THE CRITERION OF SEVENTH GRADE SUCCESS

In Chapters II and III the relationships between test scores and pupils' marks were discussed. These marks were considered as representing the pupils' ability in certain subjects. Since a pupil is admitted to a school, promoted, retarded, and sometimes even eliminated on the basis of the teacher's estimate of his work, the reliability of this estimate is of great importance.

Dearborn¹ pointed out in his investigation the large inequalities in the standards of grading employed by different teachers. Of two instructors in the same department, one gave 43% of his students the grade of "excellent" and to none the grade of "failure," whereas the other gave to none of his students the grade of "excellent," and to 14% the grade of "failure." The wide difference in this instance is no doubt due in part to a difference in the students and in the nature of the work, but largely to a difference in the standards of marking.

Starch and Elliott² endeavored to determine the range of variation and the reliability of the marks assigned by teachers to examination papers. Two examination papers written by two pupils at the end of the first year's work in English were marked by 142 teachers of different schools.

They followed this by having one geometry paper graded by 118 teachers of mathematics, and one American history examination paper by 70 history teachers. Their conclusion was that teachers differ enormously in evaluating the same pieces of work in terms of the ordinary percentage scale, and that they differ as much in one subject as another.

¹ W. F. Dearborn—School and University Grades, Bulletin of the University of Wisconsin, No. 368 (1910), page 57.

² Daniel Starch and Edward C. Elliott—Reliability of the Grading of High School Work in English, The School Review, Vol. 20 (1912), pages 442-557.

Ralph E. Carter,³ in Milwaukee, Wis., in 1911, investigated classes completing the eighth grade in 1907, and found the following variation in the marks given by the schools in arithmetic:

In School A, $\frac{1}{3}$ were below 79, and $\frac{1}{3}$ above 84

In School B, $\frac{1}{3}$ were below 71, and $\frac{1}{3}$ above 78

In School C, $\frac{1}{3}$ were below 82, and $\frac{1}{3}$ above 88

Two thirds of School B fall within the range of the lowest third of School A, while two thirds of School C fall within the range of the highest third of School B, with a margin of four points to spare. A follow up of these pupils revealed the fact that the school grading the lowest had sent a larger proportion of its poorer members to the high school than the school grading highest. Nevertheless, when all the algebra marks of the members from the three schools were ranked together, it was found that a greater percentage of the lowest school excelled in maintaining their original mark or increasing it. In fact there was a complete reversal of things from what the absolute marks alone might indicate.

Walter R. Miles⁴ made a similar study in Iowa City, and found that the average of the fifteen coefficients of correlation between one elementary school subject and another was .567; the average of the 10 between one high school subject and another was .618; the average of the 18 between elementary school subjects and high school subjects was .446. The highest coefficient was between the average of all elementary school grades and the average of all high school grades. This was .71, which does not seem very high when we remember that the marks used in these calculations were the averages of several teachers' ratings, but it is higher than is generally found by using group intelligence or achievement tests.

Frederick J. Kelly⁵ supplemented these investigations by a study of the cumulative record cards for Hackensack, N. J.,

³ Ralph E. Carter—Correlation of Elementary Schools and High Schools, *Elementary School Teacher*, 12 (1911), pages 109-118.

⁴ Walter R. Miles—Comparison of Elementary and High School Grades, *University of Iowa, Studies in Education*, Vol. 1, No. 1 (1911).

⁵ Frederick J. Kelly—Teachers' Marks, Their Variability and Standardization—*Teachers College*, 1914, Columbia University, Contributions to Education, No. 66.

schools. The term marks of 88 pupils from four 6A grades were compared with the 7B term marks of the same pupils. There was such a variation that Kelly concluded that a term mark means many different things to different teachers when rating pupils for promotion. In a study of marks given by teachers on examination papers, he contended that little significance could be placed upon uniform tests when they are marked by a single judge, or unless a rating scale is used to reduce the variations in marking.

F. W. Johnson,⁶ Principal of the University High School of Chicago, investigated the marks given by the various departments in his school for the years 1907-1908 and 1908-1909 to determine the variation among them. He found a wide variation among the various departments and a still wider one among individual teachers of the same department.

Both Franklin O. Smith,⁷ at the University of Iowa, and Ryland Newman Dempster,⁸ of Johns Hopkins, found a correlation of only .53 between high school averages and first year college averages. Dempster found, however, that the intelligence tests correlated but .47 with first year college averages. W. W. Pettit,⁹ in a comparative study of the New York High School and Columbia College Grades, found a correlation of .63 existing between the averages of the high schools and first year college marks.

Hawks¹⁰ investigated the value of high school marks for predicting first term freshman scholarship at Gettysburg College, Johns Hopkins University, and Vanderbilt University. She found a correlation of .64 between high school and fresh-

⁶ F. W. Johnson—A Study of High School Grades, *School Review*, 19: 13-24.

⁷ Franklin O. Smith—A Rational Basis for Determining Fitness for College Entrance, *University of Iowa, Studies in Education*, Vol. 1, No. 3.

⁸ Ryland Newman Dempster—Coordination of the Preparatory Training to Work in College as Based on Students' Records, Master's Essay, Johns Hopkins University, 1922.

⁹ W. W. Pettit—A Comparative Study of New York High School and Columbia College Grades, Master's Essay, Teachers College, Columbia University, 1912.

¹⁰ Lena J. Hawks—Certain Relationships Between Scholarship in High School and in College, Ph. D. Dissertation, 1929, Johns Hopkins University.

man marks at Gettysburg ; .66, .77, and .72 for three entering classes at Johns Hopkins, and .69 at Vanderbilt.

At last, however, a challenger of the teacher's mark appeared. In "Education," September, 1927, Frederick E. Bolton,¹¹ of the University of Washington, Seattle, reviewed Starch's experiment in which 10 University of Wisconsin instructors graded 10 examination papers in English, and using as a criterion of correct marking the average of the ratings given by 10 instructors, found the deviation of each instructor's mark from the average, and found that the variability of marks assigned to a single set of papers was not so great as Starch found when he took the range from the lowest to the highest mark. Bolton's procedure is better than Starch's in this respect; that he used the average as a criterion of the correct mark for a paper, and then found the variability of each teacher's mark from the probably correct mark.

E. J. Ashbaugh,¹² at the Ohio State University, had mimeographed copies of a type paper on the Stone Reasoning Test in Arithmetic marked three times at four week intervals by 49 graduate students and college seniors preparing to teach. A distribution of the total scores at the end of the first marking revealed a total range of 51; at the end of the second 50; at the end of the third 39. After a class discussion of the various problems, the students came to an agreement on what was to be marked and of the relative value to be assigned to the various elements. After that a fourth rating revealed a distribution of only 18 points.

These critical studies of the fallibility of teachers' marks may have portrayed the real facts of the situation. However, all of these investigations deal with marks given by teachers to a set of papers. The unreliability of such marks must not be wrongly confused with that of quarterly marks, with which this study is concerned.

By reliability is meant the consistency with which a test or examination or teacher's mark measures what it really does

¹¹ Frederick E. Bolton—Do Teachers' Marks Vary as Much as Supposed?—*Education* 48 (September, 1927), pages 23-39.

¹² E. J. Ashbaugh—Reducing the Variability in Teachers' Marks, *Journal of Educational Research*, Vol. 9 (1924), pages 185-198.

measure, not what it purports to measure. Reliability is synonymous with accuracy of measurement. All studies previously mentioned have endeavored to show the reliability of examination marks or term marks by comparing the marks given by one teacher on a paper or to a pupil with those given by another teacher. Such comparisons show variations in the teachers' judgments. This investigation, however, is not concerned with the difference of teachers' opinions about a paper in which they may or may not be interested, but primarily in how well a teacher's marks measure what he has taught, as represented by the recitations, written lessons, tests, and the like, demanded of his pupils. Therefore, in this study no comparisons are made between the marks given by various teachers, but a thorough analysis of the term marks of several teachers is made, with the purpose of finding out the reliability of each element that has any bearing on the final term mark given the pupil. By determining the validity of each element concerned, the reliability of the final grade can be determined.

For this study 226 pupils, the entering 7B class of September, 1929, at the Edgar Allan Poe Junior High School, were selected. These pupils were given the Otis Group Intelligence Test, the Stanford Arithmetic Test, and the Sangren Woody Reading Test. The pupils were grouped in seven classes according to the grades made in the elementary school and the results of these tests. Five teachers of English, two of mathematics, one of geography, and three of history, to whom these pupils were assigned, were asked to present their 7B work as usual, and to keep every written lesson, test, and paper of any kind whose mark was taken into consideration when making the quarterly average. At the end of the semester a careful checking and study of every type of mark and the work for which it was given gave some interesting information relative to the reliability of the teachers' term marks.

RELIABILITY OF ENGLISH MARKS

The English marks are considered first.

Teacher "A" taught two 7B classes composed of 78 very good students. Each pupil's average for the first quarter was

based on the marks received on six written compositions, three oral recitations, one test on "Asgard Stories," one grammar test, one spelling test, and one note book mark. The six written composition marks and the three oral recitation marks were averaged to count as one third of the final score; the literature test, the grammar test, and the spelling test composed the second third; the mark received on the note book counted as the last third. An average of these three gave the mark for the first quarter.

The first step was to find the reliability of the composition marks. This was done by having one set of compositions marked twice, one week apart, placing the marks on separate paper rather than on the compositions, and correlating the two sets of marks. The coefficient of reliability was .702. This, then, was taken to represent the reliability of the mark of each written composition marked during the first quarter. The next step was to determine the reliability of the three marks given for oral work. This was done by correlating the first set of marks with the second.

Apparently there was little relationship here, as the reliability coefficient came out to only .133. Counting the reliability of each of the six compositions as .702, and of each recitation as .133, then combining these coefficients by the formula,

$$rpq = \left(\frac{\frac{Nr^1 + r^2 + r^3 + r^4 + r^5 + r^6 + r^7 + r^8 + r^9}{n}}{1 + (n-1)r} \right)$$

derived from the Spearman-Brown formula,

$$rnn = \frac{nr}{1 + (n-1)r}$$

a reliability coefficient of .906 was obtained for the first third of the quarterly mark.

To determine the coefficient of reliability of the second third of the term mark, it was necessary to take the test on the Asgard Stories, the grammar test, and the spelling test, break each one into chance halves, correlate the half scores of each and then "step up" the obtained "*r*" by the Spearman-Brown formula so as to approximate the "*r*" which would

have been obtained by correlating similar forms. For the literature test, which was composed largely of essay type questions, the " r " was .466. For the grammar test it was .774. In spelling the majority of the pupils made perfect scores, and consequently the obtained " r " was very low. By combining the " r 's" of the literature, the grammar, and the spelling tests, an " r " of .797 was obtained for the second third of the quarterly mark.

The English note books, like the compositions, were marked one week apart, and then the marks were correlated. Here the " r " was .667.

By combining these three coefficients of reliability (.905, .797, and .667) a coefficient of reliability of .920 was obtained for the first quarter. (See Table 16.)

The second quarter's mark was obtained in a similar manner. Six written compositions and three oral marks counted as one third; a test on Treasure Island, a test on the Christmas Carol, and a spelling test counted as another third, and the note book counted as the rest. No marks for grammar were given, probably because it was not necessary to teach it separately to these classes of good students. To find the effect of a longer intervening period, the compositions and note books were marked three weeks apart instead of one. As a guard against possible subjectivity, the pupils' names were cut from the papers before they were marked, and numbers were substituted. The longer intervening period apparently had some effect, as the correlation for the compositions dropped from .702 to .664, and for note books from .667 to .421. The class room marks were lower still, showing an " r " of only .071. Since many factors enter into a pupil's oral recitations, the coefficients of correlation here probably mean little or nothing. The literature and spelling tests were again broken into halves and correlated. The Treasure Island test alone showed a fair correlation, .707, while the Christmas Carol and spelling showed .414 and .463, respectively.

A study of Table 16 will show that no single element of the teachers' quarterly mark was a very reliable measure. Yet the final grades, because they were made up of several marks rather than a very few, show a rather high degree of reliability.

TABLE 16
SHOWING THE COEFFICIENTS OF RELIABILITY OF THE VARIOUS
ELEMENTS AND OF THE QUARTERLY MARKS GIVEN BY
TEACHER "A" TO 78 PUPILS IN 7B ENGLISH

FIRST QUARTER			N=78
1. Written Compositions (marked one week apart)	.702	.906	.920
2. Written Compositions (marked one week apart)	.702		
3. Written Compositions (marked one week apart)	.702		
4. Written Compositions (marked one week apart)	.702		
5. Written Compositions (marked one week apart)	.702		
6. Written Compositions (marked one week apart)	.702		
7. Class Room Mark (oral)	.133	.797	
8. Class Room Mark (oral)	.133		
9. Class Room Mark (oral)	.133		
10. Literature Test (Asgard Stories)	.466	.667	
11. Grammar Test	.774		
12. Spelling Test	.463		
13. Note Book (marked one week apart)	.667	.667	
SECOND QUARTER			
1. Written Compositions (marked three weeks apart)	.664	.888	.865
2. Written Compositions (marked three weeks apart)	.664		
3. Written Compositions (marked three weeks apart)	.664		
4. Written Compositions (marked three weeks apart)	.664		
5. Written Compositions (marked three weeks apart)	.664		
6. Written Compositions (marked three weeks apart)	.664		
7. Class Room Mark (oral)	.071	.735	
8. Class Room Mark (oral)	.071		
9. Class Room Mark (oral)	.071		
10. Literature Test (Treasure Island)	.707	.421	
11. Literature Test (Christmas Carol)	.414		
12. Spelling Test			
13. Note Book (marked three weeks apart)	.421	.421	

Teacher "B" taught two classes composed of 59 pupils of average ability. Each pupil's average for the first quarter was based on the following: two grammar tests, three written compositions, one note book mark, one mark based on how well and how often he did his home work, one mark for his oral work, one mark for reading and literature, and one mark for spelling. By breaking the grammar tests into halves and correlating, r 's of .563 and .889 were obtained. The compositions, graded one week apart, revealed a surprisingly high degree of constancy in marking with an " r " of .961. The note books, marked in the same manner, came out somewhat lower, with an " r " of .785. The marks given for the home work, oral work, reading and literature, and spelling registered simply the teachers' opinions. Without a doubt they had some degree of reliability, but in working out the final coefficient, each mark was counted as zero, so that the coefficient obtained would be the lowest possible estimate of the reliability of the mark in question. In spite of this severe treatment, Table 17 shows a coefficient of reliability of .914.

The average for the second quarter was obtained from the following: four grammar tests, one note book mark, one spelling mark, one home work mark, one reading and literature mark, one grade for oral work, and one composition mark, based on the scores made on these compositions. The four grammar tests revealed reliability coefficients of .782, .799, .798, and .748. During this quarter the teacher forgot to give the second marking to either the compositions or note books, and, therefore, the coefficients of the first quarter were used. Again counting the reliability of the spelling, home work, reading and literature, and oral work marks as zero, Table 17 shows a quarterly reliability coefficient of .905.

Although for both quarters four marks based on the teacher's estimates were counted as zero, the final coefficients of reliability are high. If we admit there must be some reliability in the teacher's estimate, the final coefficients become very high indeed.

Teacher "C" taught one class of 29 pupils who were slightly below the average in ability. Her first quarterly mark

was based on the grades given on five compositions, one literature test, three grammar tests, one note book, one spelling test, and one mark based on 20 class room marks. The compositions, treated as before, yielded a reliability quotient of .934; the literature test split into chance halves and correlated, .739; the grammar tests, .279, .831, and .342; the note book, .924; the spelling test, .565; the daily recitations mark, based on

TABLE 17

SHOWING THE COEFFICIENTS OF RELIABILITY OF THE QUARTERLY ENGLISH MARKS, TOGETHER WITH THE RELIABILITY OF EACH FACTOR OF WHICH THEY ARE COMPOSED, GIVEN BY TEACHER "B" TO 59 7B PUPILS

FIRST QUARTER		N=59
1. Grammar Test	.563	.914
2. Grammar Test	.889	
3. Composition	.961	
4. Composition	.961	
5. Composition	.961	
6. Note Books (marked one week apart)	.785	
7. Home Work	?	
8. Oral Work	?	
9. Reading and Literature	?	
10. Spelling	?	
SECOND QUARTER		
1. Grammar Test	.782	.905
2. Grammar Test	.799	
3. Grammar Test	.798	
4. Grammar Test	.748	
5. Note Books	.785	
6. Spelling	?	
7. Home Work	?	
8. Reading and Literature	?	
9. Oral Work	?	
10. Composition A	.961	.986
Composition B	.961	
Composition C	.961	

20 marks for each pupil, .745. These when combined, give an exceedingly high reliability coefficient of .974. (See Table 18.)

The second quarter's average was based on the marks of five compositions, two spelling tests, four grammar tests, one note book, one literature test, and one mark obtained by averaging 27 daily recitations marks. In this case, as with Teacher "A," the compositions and note books of the second quarter were marked three weeks apart. This time the "*r*" for the two markings of the compositions dropped to .523, but for

the note books it reached .973, slightly higher than before. The spelling tests, too, correlated poorly, with an "*r*" of .565 for each.

The grammar tests likewise, when broken into chance halves and correlated, revealed the low reliability quotients of .498, .572, .451, and .579. The literature test was in a similar classification with an "*r*" of .497. The mark based on 27 daily

TABLE 18
SHOWING THE COEFFICIENT OF RELIABILITY OF TEACHER "C's"
QUARTERLY 7B ENGLISH MARKS, TOGETHER WITH THE
RELIABILITY OF EACH MARK OF WHICH THEY ARE COMPOSED

FIRST QUARTER		N = 29
1. Written Composition (marked one week apart)	.934	.974
2. Written Composition (marked one week apart)	.934	
3. Written Composition (marked one week apart)	.934	
4. Written Composition (marked one week apart)	.934	
5. Written Composition (marked one week apart)	.934	
6. Literature Test (Treasure Island)	.739	
7. Grammar Test	.279	
8. Grammar Test	.831	
9. Grammar Test	.342	
10. Note Books (marked one week apart)	.924	
11. Spelling Test	.565	
12. Daily Recitations (20)	.745	
SECOND QUARTER		
1. Written Composition (marked two weeks apart)	.523	.950
2. Written Composition (marked two weeks apart)	.523	
3. Written Composition (marked two weeks apart)	.523	
4. Written Composition (marked two weeks apart)	.523	
5. Written Composition (marked two weeks apart)	.523	
6. Spelling Test	.565	
7. Spelling Test	.565	
8. Grammar Test	.498	
9. Grammar Test	.572	
10. Grammar Test	.451	
11. Grammar Test	.579	
12. Note Books (marked three weeks apart)	.973	
13. Literature Test (Asgard Stories)	.497	
14. Daily Recitations (27)	.776	

recitations showed .776, an outstanding contrast to those of other teachers based on two or three. (See Table 18.)

The case of Teacher "C," like that of Teacher "A," shows the value of many marks. With the exception of the grades for the note books and daily recitations, the coefficients of reliability reveal the individual marks as rather inadequate measures of the pupils' worth, but the combination of many indicates an accurate measure.

Teacher "D" taught English to a class of 31 pupils of low ability. Her first quarter's marks were compiled in a rather complicated way. The marks of six spelling exercises were averaged, two spelling tests were averaged, and the average of these two averages gave the pupils' term marks in spelling.

TABLE 19

SHOWING THE RELIABILITY OF TEACHER "D'S" QUARTERLY 7B
ENGLISH MARKS, TOGETHER WITH THE RELIABILITY OF EACH
MARK OF WHICH THEY ARE COMPOSED

FIRST QUARTER				N=31
1. Spelling (6 exercises)	.965	}	.858	}
2. Spelling Tests (2)	.858			
3. Grammar Test	.744	}	.857	}
4. Grammar Test	.737			
5. Grammar Class Mark (opinion)	?	}	0	}
6. Composition	.801	}	.953	}
7. Composition	.801			
8. Composition	.801			
9. Composition	.801			
10. Composition	.801			
11. Class Mark (opinion)	?	}	0	}
12. Note Books	.791			
			.791	.906
SECOND QUARTER				
1. Spelling Test	.919	}	.958	}
2. Spelling Test	.919			
3. Grammar Test	.571	}	.696	}
4. Grammar Test	.701			
5. Grammar (class work)	?	}		}
6. Composition	.766	}	.928	}
7. Composition	.766			
8. Composition	.766			
9. Composition	.766			
10. Literature (opinion)	?		0	.879

In grammar a class room mark based entirely on the teacher's estimate was taken with the average of two grammar tests to give the final mark. Five compositions were averaged for the composition mark. There was a general class room mark based entirely on the teacher's estimate, and one mark for the note book. Then these five were averaged to secure the quarterly mark. (See Table 19.)

The reliability of the mark for the six spelling exercises was obtained by correlating three of the exercises with the other three. The two spelling tests correlated .858. By combining these a coefficient of reliability of .958 was obtained for the spelling. The two grammar tests correlated .744 and .737, making .857 for the two. As was done before, the teacher's opinion of the class room work in grammar was counted as zero. This zero, averaged with the "*r*" of .857, gave a reliability coefficient of only .593 for the grammar. The second numbering of the compositions, one week after the first, gave a reliability of .801. When taken for five compositions, this became .953. The teacher's opinion of the class room work was counted as zero, and the correlation of the first and second note book marks was .791. Combining the coefficients of reliability for the averages of spelling, grammar, written compositions, class room work, and note books, we have another high coefficient of .906. (See Table 19.)

For the second quarter two spelling tests were averaged for the spelling mark, two grammar tests and one teacher's estimate for the grammar, four written compositions for the composition mark, and one teacher's opinion for the grade in literature.

The two spelling tests, correlating .919 and .919, gave an average reliability of .958. The two grammar tests gave "*r*'s" of .571, and .701; these, when combined with an assumed zero for class recitations, gave an "*r*" of .696. Teacher "D," like the others, seemed affected by the three weeks' interval between the composition markings, and dropped her coefficient of reliability on the marking from .801 to .766. For the four together, however, an "*r*" of .928 was obtained. The teachers' estimate for the literature work was again counted as zero.

Here, as before, the combination of several numbers to get the term average gives a final mark with a high reliability coefficient. Again assuming that the teacher's opinions have some validity, the final coefficient of correlation would become still higher.

Teacher "E" taught English to 29 very slow pupils, most of whom were taking the work for the second time. For each quarterly mark the average of four compositions was taken as one fourth; a mark for reading, based entirely on the teacher's estimate, was the second; a grammar mark, based on a sort of general estimate of how well the pupil mastered each topic, was the third; an estimate of how well and how often

TABLE 20

SHOWING THE COEFFICIENT OF RELIABILITY OF TEACHER "E's"
QUARTERLY 7B ENGLISH MARKS, TOGETHER WITH THE
RELIABILITY OF EACH MARK OF WHICH THEY ARE COMPOSED

			N=29	
FIRST QUARTER				
1. Written Composition (marked one week apart)	.832	.946	.756	
2. Written Composition (marked one week apart)	.832			
3. Written Composition (marked one week apart)	.832			
4. Written Composition (marked one week apart)	.832			
5. Reading	?	0		
6. Grammar	.804	.804		
7. Home Work	?	0		
SECOND QUARTER				
1. Written Composition (marked three weeks apart)	.750	.923	.752	
2. Written Composition (marked three weeks apart)	.750			
3. Written Composition (marked three weeks apart)	.750			
4. Written Composition (marked three weeks apart)	.750			
5. Reading	?	0		
6. Grammar	.804	.804		
7. Home Work	?	0		

the pupil did his home work was the fourth. A second marking of the written composition in the first quarter, taking place one week after the first marking, and three weeks after the first marking in the second quarter, gave reliability coefficients of .832, and .750, respectively. For the four compositions these "stepped up" to .946 for the first quarter and .923 for the second. For the reading and home work estimate, which could not be measured, the reliability was taken as zero. (See Table 20.)

For grammar, however, the teacher had kept an apparently accurate but unusual record.* A portion of it is given below:

Pupil	Pronoun	Verb	Object	Adjective	Adverb	Average
a	I	I	2	IS	I	70
b	I	I½	I	I	I	75
c		Iu	I	Iu		45
d		IS	I	IS	I	60
e		IS	I	2S	I	80
f				I	I	35
g		I				60
h	I	IS	I	IS	2	70
i		IS	I	IS	I	70
j		I	I	I	I	60

At the end of the time allotted to each topic in grammar, if a pupil had mastered a topic to any extent he was marked with a "I." If he had barely passed, a "u" was placed after the mark: If he had done the work unusually well, an "S" was added. At the end of the quarter the teacher's final estimate was based on her reading of these marks. To test the reliability of this, the column in which the marks were given was cut out from the record sheet and two months later the same sheet was returned to her for a second marking. In spite of the long period intervening, the correlation of the two markings revealed a coefficient of reliability of .804. Counting the marks in reading and home work as having a reliability of zero, the coefficient of reliability for the first quarter came out to .756 and for the second quarter .752. It is quite probable that both the teacher's reading estimates and her home work estimates had some definite validity, and therefore the chances are that the final mark had a much higher coefficient of reliability than Table 20 indicates. Apparently the marking of the grammar with symbols rather than numbers did not lower the reliability.

A review of Tables 16, 17, 18, 19, and 20 will show a fairly high degree of reliability in the quarterly marks of these five teachers.

RELIABILITY OF MATHEMATICS MARKS

The mathematics of the pupils involved in this study were taught by two teachers, Teacher "F," having one class of 29 very bright pupils, and Teacher "G," having the remaining 187 in six classes. The quarterly mark of Teacher "F" was

the arithmetical average of four tests and one note book mark. Each test was divided into chance halves and then correlated and "stepped up" by the Spearman-Brown formula.

Test One, on Geometric Forms, gave a reliability coefficient of .444; Test Two, on Lines, a reliability of .419; Test Three, on Graphs, gave a reliability of .456, and Test Four, on Angles, gave a reliability of .726. Two markings of the note books were made, the first on November 8th and the second on November 15th. Although the marks apparently did not vary very much, generally not more than five points, the coefficient of reliability turned out to be $-.072$. When these

TABLE 21
SHOWING THE COEFFICIENT RELIABILITY OF TEACHER "F's"
QUARTERLY 7B MARKS IN MATHEMATICS, TOGETHER WITH THE
RELIABILITY OF EACH MARK OF WHICH THEY ARE COMPOSED

FIRST QUARTER		N = 29
1. Geometric Forms	.444	} .765
2. Lines	.419	
3. Graphs	.456	
4. Angles	.726	
5. Note Books	— .072	
SECOND QUARTER		
1. Test One	.738	} .867
2. Test Two	.665	
3. Test Three	.374	
4. Test Four	.321	
5. Test Five	.744	
6. Test Six	.625	
7. Note Books	— .072	

five were put together, the reliability coefficient for the first quarter was determined at .765. (See Table 21.)

For the second quarter the term average was gotten from six tests and one note book mark. The reliability of the first test was .738; of the second, .665; of the third, .374; of the fourth, .321; of the fifth, .744, and of the sixth, .625. Since the teacher neglected to give the second marking of the note book during the second quarter, the coefficient of $-.072$ received from the marking of the first quarter's note books was used for the second quarter. All of these taken together gave a final coefficient of reliability of .867.

Teacher "G," who taught 187 pupils, based her quarterly marks on three tests. For the first quarter Test One, on Geometric Figures, gave a coefficient of reliability of .746; Test Two, on Lines, a reliability of .788, and Test Three, on Curves, a reliability of .645. The final reliability coefficient for the quarter was .850.

For the second quarter Test One, on Angles, gave a reliability of .447; Test Two on Geometric Construction, a reliability of .562, and Test Three, a reliability of .201. This gave the quarterly mark a reliability of .670. (See Table 22.)

It will be noticed that the reliability of the various tests in this subject is low. The 7B mathematics is principally made

TABLE 22
SHOWING THE COEFFICIENTS OF RELIABILITY OF TEACHER "G's"
QUARTERLY 7B MARKS IN MATHEMATICS, TOGETHER WITH
THE RELIABILITY OF EACH MARK OF WHICH THEY
ARE COMPOSED

FIRST QUARTER		N=187
1. Geometric Forms	.746	}.850
2. Lines	.788	
3. Graphs	.645	
SECOND QUARTER		
1. Angles	.447	}.670
2. Test Two	.201	
3. Geometric Construction	.562	

up of intuitive geometry, but the arithmetic of the former grades is constantly involved. Apparently most mathematics tests contain only a few questions, and are not a reliable means of determining the pupils' ability in the subject. Consequently the tests should either be lengthened or else given more frequently. A larger number of tests by either teacher would have greatly increased the reliability coefficient of the quarterly mark. Probably each test should include a larger number of items that could be answered quickly. This is done in most standardized tests. If the test could not be administered in a single period, use of more time is justifiable if the test has a higher reliability. The Columbia Research Bureau Algebra Test by Otis and Wood, including forty items, has a reliability of .86.

RELIABILITY OF GEOGRAPHY MARKS

Teacher "F," the only geography teacher involved in this study, taught 226 pupils divided among seven classes. No class room marks were given, and every test, with the exception of the first one on Definitions, had been made up under the direction of the Supervisor of Geography in the Junior High Schools. These tests included from 20 to 50 objective questions in each. There was almost no opportunity for the

TABLE 23

SHOWING THE COEFFICIENT RELIABILITY OF TEACHER "H's"
QUARTERLY 7B MARKS IN GEOGRAPHY, TOGETHER WITH
THE RELIABILITY OF EACH MARK OF WHICH
THEY ARE COMPOSED

N=221

FIRST QUARTER			
1. Definitions Test	.416	} .913	} .951
2. Wheat Test	.834		
3. Corn Test	.733		
4. Rice Test	.970		
5. Rotation Test	.878	} .882	
6. Achievement Test	.700		
7. Note Books (marked one week apart)	.828	.828	
SECOND QUARTER			
1. Fish Practice Test	.857	.857	} .960
2. Meat Achievement	.774	} .880	
3. Seafood Achievement	.688		
4. Fruit Achievement	.666		
5. Note Books (marked three weeks apart)	.933	.933	

teacher's opinion to enter into the marking. The marking of the note books at the end of each quarter was the only instance in which the teacher's individuality could enter into the final mark.

For the first quarter Test One, on Definitions, Test Two, on Wheat, Test Three on Corn, and Test Four, on Rice, were averaged for one third of the final mark. The coefficients of these four were .416, .834, .733, and .970, respectively. (See Table 23.)

A fifth test, known as the Rotation Test, with a reliability coefficient of .878, and a sixth test giving a final review of all the others, and having a reliability of .700, were averaged for the second third of the term mark. The note books were taken for the last third. Here, as in other cases, the note books were marked a second time after an interval of one week. The coefficient of reliability was .828. A reliability coefficient for the first four tests was .913; for the second group of two, .882; and for the note book .828. These gave the quarterly mark a reliability of .951.

During the second quarter a practice test on Fish was taken as one third of the quarterly mark. This had a reliability of .857. Three other tests on Meat, Seafood, and Fruit were averaged for the second third of the quarterly mark. Their coefficients of reliability were .774, .688, and .666, respectively, giving a final reliability of .880. The note books, which counted as the third part of the quarterly mark, were marked the second time after an interval of three weeks and showed a reliability coefficient of .933. These gave the term mark of the second quarter a reliability coefficient of .960. (See Table 23.)

This study makes no statement regarding the advisability of giving the same test to all classes throughout the school and the city; however, it shows that when a quarterly or a term mark is determined almost solely on the basis of several objective tests, we have about as reliable measure of what the pupil has done as can be obtained.

RELIABILITY OF HISTORY MARKS

The history was taught the pupils involved in this study by three teachers. Teacher "I" taught one class of 39 pupils of high ability; Teacher "J" taught five classes, composed of 138 pupils of varying abilities, and Teacher "K," one class made up of 29 slow pupils, most of whom were repeating the term's work.

Teacher "I" for the first quarter based her average on the following:

1. The average of six class room marks.
2. One note book mark.
3. A test given about the middle of the quarter.
4. A test given at the end of the quarter.

The class room marks, when divided and correlated, gave a reliability coefficient of .908 for their average. The note books, marked one week apart, gave a reliability coefficient of .743. The two tests revealed reliability measures of .710 and .929.

TABLE 24

SHOWING THE COEFFICIENT OF RELIABILITY OF TEACHER "I's"
QUARTERLY 7B MARKS IN HISTORY, TOGETHER WITH
THE RELIABILITY OF EACH MARK OF WHICH
THEY ARE COMPOSED

FIRST QUARTER		N=39
1. Class Room Marks (6 marks)	.908	} .949
2. Note Books	.743	
3. Test	.710	
4. Test	.929	
SECOND QUARTER		
1. Class Room Marks (6 marks)	.929	} .904
2. Note Books	.578	
3. Test	.541	
4. Test	.469	
5. Test	.751	

.929, respectively. This made the quarterly reliability coefficient .949.

For the second quarter her average was obtained as for the first, except a third test was given and the note books were marked three weeks apart instead of two. Here the coefficient of reliability for the six class room marks was .929; for the note books .578; for Test One .541; for Test Two .469, and for Test Three .751. When taken together these five marks gave the quarterly average a coefficient or reliability of .904. (See Table 24.)

Teacher "J," who had 138 pupils, kept no class room marks of any kind, but based the quarterly average of the pupils on

two tests and a third mark based on his opinion of the worth of the pupil. This third mark could have been influenced by the bringing in of additional work, a display of extra effort, any evidence of interest or lack of interest on the part of the pupil, or any other factors that might have been brought to the teacher's attention. While notes indicating such things were in some cases inserted in the teacher's record book, in most cases there was nothing.

For the first quarter Test One gave a reliability coefficient of .666, and Test Two a reliability of .600. The class mark

TABLE 25
SHOWING THE COEFFICIENT OF RELIABILITY OF TEACHER "J's"
QUARTERLY 7B MARKS IN HISTORY, TOGETHER WITH THE
RELIABILITY OF EACH MARK OF WHICH
THEY ARE COMPOSED

FIRST QUARTER		N=138		
1. Test	.666	} *.815	†.686	
2. Test	.600			
3. Class Mark	.519			
SECOND QUARTER				
1. Test	.785	} *.928	†.874	
2. Test	.788			
3. Note Books (marked three weeks apart)	.963			
4. Class Mark	.519			
* Class Room work counted as .519				
† Class Room work counted as 0				

of the first quarter was correlated with the class mark of the second quarter to see if there was any relationship between them. These revealed an "*r*" of .519. Perhaps this meant nothing; it was included, however, in determining the quarterly coefficient of reliability and gave an "*r*" of .815. Counting, however, the class room mark as zero, the coefficient of reliability was .686. (See Table 25.)

For the second quarter the two tests revealed reliability coefficients of .785 and .788, and the note books, marked three weeks apart, gave a reliability coefficient of .963. Again when we counted .519 as the reliability coefficient for the class mark, the "*r*" for the quarter was .928. Counting the class mark as zero, the reliability coefficient was .874.

Here, as in mathematics, had additional tests been given, the final coefficient of reliability would have been much higher. This shows the advisability of basing quarterly marks on several numbers rather than on a few.

Teacher "K" taught history to a class of 29 pupils, nearly all of whom were repeating the term's work. The mark for each quarter was based on the following:

1. Note book mark.
2. A test given about the middle of the quarter.
3. A test given at the end of the quarter.
4. A class room mark based on five short papers.

In order to give the tests more weight, the result of each was counted double in the final averaging. The note books,

TABLE 26
SHOWING THE COEFFICIENT OF RELIABILITY OF TEACHER "K's"
QUARTERLY 7B MARKS IN HISTORY, TOGETHER WITH THE
RELIABILITY OF EACH MARK OF WHICH
THEY ARE COMPOSED

N=29

FIRST QUARTER

1. Note Books (marked one week apart)	.932	} .763
2. Test (counted double)	.345	
3. Test (counted double)	.502	
4. Class Room Marks (five short papers)	.509	

SECOND QUARTER

1. Note Books (marked three weeks apart)	.806	} .921
2. Test (counted double)	.870	
3. Test (counted double)	.787	
4. Class Room Marks	.644	

marked one week apart, revealed an "r" of .932; Test One .345; Test Two .502; and the class room marks, when divided and correlated, .509. Weighting both tests as the teacher had done gave the quarterly mark a reliability coefficient of .763.

In the second quarter exactly the same procedure was followed, except the note books were marked three weeks apart instead of one. The coefficients of reliability of these were as follows: Note book .963; Test One .870; Test Two .787; and for class room marks .644. These gave a final reliability coefficient of .921. (See Table 26.)

A review of Tables 24, 25, and 26 shows, as in other cases, that while the reliability of a single mark may not be very high, yet the quarterly marks show a high degree of reliability, providing they are based on several numbers rather than on a few.

TABLE 27

SHOWING THE AVERAGE RELIABILITY OF THE QUARTERLY MARKS FOR EACH TEACHER, FOR EACH SUBJECT, AND FOR ALL TEACHERS

Subject	Number pupils	First quarter	Second quarter	Average
ENGLISH				
Teacher "A"	78	.920	.865	.892
Teacher "B"	59	.914	.905	.909
Teacher "C"	29	.974	.950	.962
Teacher "D"	31	.906	.879	.892
Teacher "E"	29	.756	.752	.754
MATHEMATICS				
Teacher "F"	39	.765	.867	.816
Teacher "G"	187	.850	.670	.760
GEOGRAPHY				
Teacher "H"	226	.951	.960	.956
HISTORY				
Teacher "I"	39	.949	.904	.926
Teacher "J"	138	.686	.874	.780
Teacher "K"	29	.763	.921	.842

AVERAGE RELIABILITY OF ALL QUARTERLY MARKS

English	.881
Mathematics	.788
Geography	.956
History	.849
All Teachers	.863

Table 27 shows in concise form the average reliability of the quarterly marks for each teacher, for each subject, and for all teachers.

According to Ruch,¹⁴ coefficients of reliability from .75 to .90 are fairly high and anything over .90 is high. A review of these quarterly coefficients of reliability shows that in English one teacher's marks had an "*r*" of over .95; three were approximately .90, and one was .75. The average for these is .881. In mathematics one was slightly over .80 and the other

¹⁴ G. M. Ruch—The Improvement of the Written Examination, 1924.

slightly over .75. In geography the only one was .95; and in history .92, .78, and .84, respectively, with an average of .849. With the exception of the marks of two teachers in mathematics, one in history, and one in English, the reliability of the quarterly marks seems to be very high. The low reliability of the marks of the English teacher is caused by the necessity of counting a reliability of zero for two of the four marks included. The chances are that those marks based on her opinion have considerable validity, and, therefore, her quarterly marks also may be quite reliable. The low coefficient of reliability for the one teacher in history and for the two in mathematics apparently was caused by not basing the quarterly mark on sufficient evidence.

TABLE 28

SHOWING THE CORRELATION BETWEEN THE SCORES MADE ON THE OTIS GROUP TEST, STANFORD AND SANGREN-WOODY READING TESTS, STANFORD ARITHMETIC TEST, AND THE AVERAGE OF MARKS FOR SEPTEMBER-FEBRUARY, 1928-1929, AND SEPTEMBER-FEBRUARY, 1929-1930

	Term mark Sept.-Feb. 1928-1929 N=166	Term mark Sept.-Feb. 1929-1930 N=221
Otis Group Test	.474	.455
Stanford Reading, 1928	.307	
Sangren-Woody, 1929		.533
Stanford Arithmetic	.556	.529

While this study was in progress it was thought that perhaps the teachers, realizing that such a check was being made on their marks, might be more careful than usual during the semester when the marks were being examined. In order to determine whether or not this was true, the scores made by the pupils in the Otis Group Test, the Sangren-Woody Reading Test, and the Stanford Arithmetic Test were correlated with the pupils' average of marks in English, mathematics, geography, and history to see if there was a higher correlation during this semester than there had been one year before when the Otis Group Test and the Stanford Tests in reading and arithmetic were given. Table 28 shows that the correlation between intelligence and the average of pupils' marks for the 1929-1930 semester was lower than that of the previous year.

The correlation between scores made in the Sangren-Woody Test and the semester averages was higher than had been the previous correlation between the Stanford Reading Test and the averages. In arithmetic there was a lower degree of correlation between the test scores and the semester average than there had been the year before. Since in the one case the correlation for the 1929-1930 semester was higher and in two instances lower than had been the case for the same semester of the previous year, apparently the study made of the teachers' marks did not increase the care with which they were given.

Since the average of the reliability of the quarterly marks of eleven teachers was .863, it would indicate that teachers' quarterly marks have a much higher degree of reliability than has usually been supposed.

CHAPTER V

CONCLUSIONS

I. In the Baltimore Junior High Schools, the Otis Group Test of Mental Ability correlates more highly with the teachers' 7B marks in English, mathematics, geography, history, or an average of these, than does the Stanford Arithmetic Test, the Stanford Reading Test, or the combination of any two, or all three of these taken together, the coefficients ranging from .278 to .541, prediction thus being but 13% better than from guessing.

II. Pupils who are at or above the seventh grade level in intelligence, reading, or arithmetic are quite likely to succeed in first year junior high school, in this study only 4% (in all schools) failing. Nine per cent of those at sixth grade level failed, 22% of those at fifth grade level, and 27% of those at third or fourth grade level.

III. In this investigation teachers' quarterly marks usually show a high degree of reliability, although there is reason to believe that the Spearman-Brown formula raises the reliability coefficients higher than they actually are.

IV. This study substantiates the statements that a teacher's mark in a single examination may be unreliable, but shows that when several such estimates are made, the reliability increases. Therefore, the teacher's term mark, when based on several marks rather than a few, is a good measure of the achievement of a pupil.

V. If a teacher marks daily recitations, a large number of such marks should be given if they are to measure accurately the work of the pupil.

VI. A quarterly or term mark should not be based on the results of only two or three tests unless they are of proven high reliability.

VII. Mathematics tests composed of only a few questions are generally unreliable. If it is impossible to lengthen the test for regular school use, several tests should be given if the final mark is to be an accurate measure of pupil achievement.

VIII. There is some evidence (see pages 37-38) which seems to indicate that numerical symbols of achievement are not necessarily more accurate than the use of letters or other symbols.

BIBLIOGRAPHY

- ASHBAUGH, E. J.: "Reducing the Variability in Teachers' Marks." *Journal of Educational Research*, 9: 185-198 (1924).
- BOLTON, FREDERICK E.: "Do Teachers' Marks Vary as Much as Supposed?" *Education*, 48: 23-39 (September, 1927).
- BOOK, W. F.: "The Intelligence of High School Scores." The Macmillan Company (1922).
- BROOKS, FOWLER D.: "The Psychology of Adolescence." Houghton, Mifflin Company (1929).
- BROOKS, FOWLER D.: "Sectioning Junior High School Pupils by Tests and School Marks." *Journal of Educational Research*, Vol. 12 (1925), pages 359-369.
- CARTER, RALPH E.: "Correlation of Elementary Schools and High Schools." *Elementary School Teacher*, Vol. 12 (1911), pages 109-118.
- DEARBORN, W. F.: "School and University Grades." *University of Wisconsin Bulletin*, No. 368 (1910).
- DEMPSTER, RYLAND NEWMAN: "Coordination of the Preparatory Training to Work in College, as Based on Students' Records." Master's Essay, Johns Hopkins University, 1922.
- GARRETT, HENRY F.: "Statistical Methods in Education and Psychology." Longmans Green and Co. (1926).
- GREGORY, C. A.: "Fundamentals of Educational Measurement." D. Appleton and Co. (1922).
- HADDOCK, NELLIE P.: "The Use of the Stanford Achievement Test to Indicate High School Success." *School and Society*, Vol. 25: 114-116 (January 22, 1927).
- HALL, E. C.: "An Experiment with the Courtis Test in Arithmetic." (Report of investigations by members of the Society of College Teachers of Education.)
- HAWKS, L. J.: "Certain Relationships Between Scholarship in High School and in College." Ph. D. Dissertation (1929), Johns Hopkins University.
- JOHNSON, F. W.: "A Comparative Study of the Grades of Pupils from Different Elementary Schools in Subjects of the First Year High School." *Elementary School Teacher*, Vol. 11 (1911), pages 63-68.
- JOHNSON, F. W.: "A Study of High School Grades." *School Review*, Vol. 19: 13-24.
- JORDAN, A. M.: "Correlation of Four Intelligence Tests with Grades." *Journal of Educational Psychology*, Vol. 13 (1922), pages 419-429.
- KELLY, JAMES K.: "Teachers' Marks—Their Variability and Standardization." *Teachers' College, Columbia University, Contributions to Education*, No. 66, 1914.
- MILLER, H. T.: "A Comparative Study of Grades of Pupils from Different Elementary Schools in Subjects of the First Year High School." *Elementary School Teacher*, Vol. 11 (1911), pages 161-175.
- OELL, CHARLES W.: "The Interpretation of the Probable Error and the Coefficient of Correlation." *University of Illinois, Bulletin* No. 32 (Aug. 31, 1926).
- OHLSON, DAVID: "School Marks Versus Intelligence Rating." *Educational Administration and Supervision*, 13: 90-102 (February, 1927).

- PETIT, W. W.: "A Comparative Study of New York High School and Columbia College Grades." Master's Essay, Teachers' College (1912).
- PROCTOR, W. M.: "Use of Psychological Tests in Educational and Vocational Guidance of High School Pupils," 1923.
- RAFFERTY, W. B.: "Graduation and Promotion in the Elementary Schools of New York." Teachers' College, Columbia University, Master's Essay (1902).
- RECTOR, W. G.: "A Study of the Prediction of High School Success." *Journal of Educational Psychology*, Vol. 16 (1925), pages 28-37.
- RICE, J. M.: "Scientific Management in Education." New York (1913).
- RIETZ, H. L., and SHANE, J.: "Correlation of Efficiency in Mathematics and Efficiency in Other Subjects." *University of Illinois, Bulletin* 6: 301-324.
- ROSS, C. C.: "The Relation Between Grade School Record and High School Achievement." *Educational Pamphlet*, Vol. 33 (1925).
- RUCH, G. M.: "The Improvement of the Written Examination." Scott, Foresman and Company (1924).
- RUGG, HAROLD A.: "Statistical Methods Applied to Education." Houghton, Mifflin Company (1917).
- RUGGLES, A. M.: "Grades and Grading." Teachers' College, Columbia University. Master's Essay, 1911.
- SENOUR, A. C.: "The Relation Between Intelligence and Success of 206 Pupils in Junior High School." Master's Thesis, University of Chicago, 1917.
- SIMPSON, B. R.: "Correlation of Mental Abilities." Teachers' College, Columbia University. *Contributions to Education*, No. 53 (1912).
- SMITH, A. G.: "Rational College Marking System." *Journal of Educational Psychology*, Vol. 1, 99: 84-96.
- SMITH, FRANKLIN O.: "A Rational Basis for Determining Fitness for College Entrance." *University of Iowa, Studies in Education*, Vol. 1, No. 3 (1910).
- STARCH, D., and ELLIOTT, E. C.: "Reliability of Grading High School Work in English." *School Review*, Vol. 20 (1912), pages 442-457.
- STARCH, D., and ELLIOTT, E. C.: "Reliability of Grading High School Work in Mathematics." *School Review*, Vol. 12 (1904), pages 254-259.
- STEELE, A. G.: "Training Teachers to Grade." *Pedagogical Seminary*, Vol. 18 (1911), pages 523ff.
- STRAYER, G. D.: "Measuring Results in Education." *Journal of Educational Psychology*, Vol. 2 (1911), pages 3-10.
- THORNDIKE, E. L.: "Theory of Educational Measurements." Teachers' College, Columbia University, 1904.
- UNZICKER, S. P.: "Teachers' Marks and Intelligence." *Journal of Educational Research*, Vol. 11 (1925), pages 123-131.
- WHITE, E. E.: "Promotion and Examinations." *U. S. Bureau of Education, Bulletin* No. 3 (1891).

INDEX

A

- ACHIEVEMENT scores, predictive value of 1, 5-8, 9-23
 ARMY-Alpha Test 5
 ASHBAUGH, E. J. 27

B

- BOLTON, F. E., variability of teachers' marks 27
 BROOKS, F. E., Psychology of Adolescence 1; sectioning of junior high school pupils 5

C

- CARTER, correlation of elementary and high school grades 25
 CHRONOLOGICAL age, predictive value of 1
 CLASSIFICATION of pupils 4, 5; basis of 1, 2
 COLLEGE entrance, basis of determining fitness for 25
 COMPARISON of elementary and high school grades 26
 COORDINATION of preparatory training and college grades 26
 CORRELATION of elementary and high school grades 25
 CORRELATION of intelligence tests and grades 5

D

- DEARBORN, W. F., school and university grades 2, 24
 DEMPSTER, R. N., preparatory training and college work 26
 DISTRIBUTION of averages of seventh grade pupils 18-23

E

- EDGAR Allan Poe Junior High School 6, 9, 10, 12; distribution of pupils in 10
 ELLIOT, STARCH AND, reliability of high school marks 2
 ENGLISH marks, reliability of 28-38
 ERROR of prediction, Kelly 8
 ESTIMATES of teachers, predictive value of 1

G

- GEOGRAPHY marks, reliability of 41, 42, 46
 GRADES, college 26; high school 26
 GROUPING, by tests and school marks, Brooks 4, 5

H

- HAWKS, L. J., Stanford Achievement Tests and high school success 6; relationship between high school scholarship and college marks 26
 HIGH School grades 26
 HISTORY marks, reliability of 42-46

I

- INTELLIGENCE scores, predictive value of 1

J

- JOHNSON, F. W., A Study of High School Grades 26
 JORDAN, correlation of intelligence tests and grades 5
 JUNIOR High School, Edgar Allan Poe 3; Number Forty 2

K

- KELLY, Error of prediction 8; teachers' marks, variability and standardization 25

M

- MARKS, average 1, 49; daily recitation 49; mathematics 38-40, 46; reading 27; reducing reliability of 27; reliability of 3, 49
 MENTAL age 4
 MILES, comparison of elementary and high school grades 25
 MILLER, Group Test 5

O

- OTIS, Group Test, correlation of teachers' marks 47, 49; predictive value 3, 5, 6-8, 9-23, 28, 49
 OTIS AND WOOD, Algebra Test 40-46

P

- PETTIT, W. W., a comparative study of high school and college grades 26
 PUPILS, success of, low ability 12-16, 18-23; average ability 16, 18-23; high ability 17, 18-23

R

- RECITATION, daily 49
 RECORDS of students 26
 RELIABILITY of English marks 28-38
 Geography marks 41, 42, 46
 High school marks 2
 History marks 42-46
 Mathematics marks 2, 38-40, 46, 49, 50
 RUCH, coefficient of reliability, 46

S

- SANGREN WOODY Reading Test 4, 47, 48

SCHOLARSHIP, high school and college 26

SCHOOL and university grades 2, 24

SMITH, F. O., basis for determining fitness for college entrance 26

SPEARMAN-BROWN Formula 29, 49

STANFORD Achievement Test, predictive value of 3, 6

STANFORD Arithmetic Test 3; correlation with teachers' marks 47, 48, 49; predictive value of 5-8, 9-23

STARCH AND ELLIOT, reliability of grades 2, 24

STONE Reasoning Test 27

SYMBOLS, accuracy of 50; use of 38

T

TEACHERS' estimates 1, 4

TEACHERS' marks, reliability of 24, 25, 49; standardization of 25; variability of 24, 25, 27

TERMAN Group Test 5

W

WOOD, Otis AND, algebra test 40, 46

VITA

The author was born in New Windsor, Carroll County, Maryland, on April 22, 1896. He attended the County Schools there until he was twelve years of age, when he was placed in the Baltimore Public Schools. He graduated from the City College in 1915, and during the following year worked for the Baltimore and Ohio Railroad Company. In the fall of 1916 he began teaching in the Evening High School at Baltimore City College and entered the Baltimore Teachers' Training School, from which he graduated in February, 1918, being permitted to complete his work in less than the usual time in order that he might enter military service in the World War. On his return from France in the Spring of 1919, he continued his work in the Evening High School and began teaching in the Public Schools of Baltimore. He was appointed teacher in Elementary School No. 71 in May, 1919, teacher in Junior High School No. 47 in September, 1919, Special Assistant in Junior High School No. 47 in January, 1922, received his degree of Bachelor of Science at Johns Hopkins University in February, 1922, was appointed Principal of the Junior High Summer School in June, 1926, and Principal of Junior High School No. 1 in September, 1926. In June, 1927, he received the degree of Master of Arts from The Johns Hopkins University.

